IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.13, No.1, January 2019, pp. 11~20
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.28076

Prediction of Length of Study of Student Applicants Using Case Based Reasoning

Ulfi Saidata Aesyi*[1], Retantyo Wardoyo[2]
[1]Department of Information Systems, FTTI UNJANI, Yogyakarta, Indonesia
[2]Department of Computer Science and Electronics, FMIPA UGM, Yogyakarta, Indonesia
e-mail: ***[1]ulfiaesyi@gmail.com**, [2]rw@ugm.ac.id

Abstrak

Kelulusan seorang mahasiswa sangat penting untuk perguruan tinggi terutama pihak program studi. Lama studi mempengaruhi nilai akrditasi dari program studi. Salah satu penilaian pada Akreditasi Program Studi Magister Buku V Pedoman Penelitian Instrumen Akreditasi standar 3 adalah profil lulusan yang mencangkup rata-rata masa studi lulusan (dalam tahun) dan rata-rata IPK lulusan.

Pada penelitian ini, sistem yang dibangun adalah sistem untuk memprediksi kelulusan calon mahasiswa S2 Ilmu Komputer Universitas Gadjah Mada. Sistem yang dibangun bertujuan untuk membantu prediksi kelulusan calon mahasiswa sehingga pihak program studi dapat menyeleksi calon mahasiswa yang akan diterima. Case Based reasoning digunakan untuk membantu proses prediksi. Prediksi dilakukan dengan memasukkan kasus baru yang berisi 13 fitur. Proses selanjutnya adalah menghitung similarities lokal dengan menggunakan Euclidean Distance Dan Hamming Distance Dan menghitung similaritas global menggunakan nearest neighbor. Hasil dari perhitungan similaritas global tertinggi akan diambil solusinya. Proses revisi dilakukan jika nilai kurang dari threshold.

Hasil penelitian menunjukkan sistem ini membantu pihak program studi untuk proses penyelenggaraan pendidikan. Hasil pengujian terhadap 50 data adalah 76%.

Kata kunci*— lama studi, Case Based Reasoning, Euclidean Distance, Hamming Distance, Nearest Neighbor.*

Abstract

Graduation is important matter in college. Length of study can be used to evaluate curriculum. It affect accreditation score of the sutdy program. Based on Akreditasi Program Studi Magister Buku V Pedoman Penilaian Instrumen Akreditasi 3rd standard there is rule about students and graduation, such as profile of the graduates including average length of study time and gpa (grade point average) of graduates.

In this study, system built to predict Gadjah Mada University Master of Computer Science student applicant's length of study. It used new case with 13 features from applicant that will be predict as new case, then calculate local similarity using euclidean distance and hamming distance while global similarity using nearest neighbor. Maximum value of global similarity taken as solution while revised will be done if it's value below threshold.

Result of this study show that system can help study program to manage educational process. It show 76% accuracy of 50 data.

Keywords*— lenght of studi, Case Based Reasoning, Euclidean Distance, Hamming Distance, Nearest Neighbor.*

Received September 5th,2017; Revised January 3rd, 2018; Accepted January 23th, 2019

1. INTRODUCTION

Graduation time of a student is very important because it deals with many parties, in addition to the student concerned, the guardian lecturer, and the head of the study program as well as related parties during the student graduation. In the Akreditasi Program Studi Magister Buku V Pedoman Penilaian Instrumen Akreditasi 3rd standard, about students and graduation. One element of the assessment is the graduate profile. The graduate profile consists of the average graduate study period (in years) and the average graduate GPA. The study period of graduates is an important thing that needs to be considered by every college because it affects the accreditation assessment. Therefore, a system is needed to predict the time of student graduation to early know the length of study of the students.

The prediction of the early graduation time of the students can help the education process for the study program. [1] revealed that the velocity of the study period was a determinant of a student taking a bachelor's degree. In this study discussed the application to predict the velocity of study of the students of State Islamic University (UIN) Syarif Hidayatullah Jakarta. This application using Cross Industry Standard Process for Data Mining (CRIPS-DM) and the algorithm to be implement is Naïve Bayes for data Clasification.

The student length of study is one of the important parameters in evaluating the student study performance, it is very reasonable if the prediction of the student length of study is needed by college management. One of the ways to make predictions by delve the data on the experiences of alumni using Case Base Reasoning (CBR). CBR is used to resolve new cases by remembering situations that have occurred by taking new solutions for similar cases. The method used in this study is CBR because CBR resolves new problems by remembering similar situations and using information and knowledge of these problems [2].

Case Based Reasoning (CBR) is a problem-solving method that uses knowledge of past experience to solve new problems [3], [4]. Cases in the past are stored by including features that describe the characteristics of the case and its solutions. There are 4 stages of the process that exist in a case-based computer reasoning system [5], there are retrieve to get similar cases, reuse using existing cases and try to solve currently problem, revise change and adopt the solution offered if its necessary, retain, use a new solution as part of the case base, then a new case is updated into the case base. The system illustration is shown in Figure 1. To improve the calculation results, [6] suggest using euclidean distance, manhattan distance, minkowski matrix, and mahalanobis distance for comparison of numerical data while for comparison of object data using hamming distance, grower-legendre, socal-michener , and jaccard similarity.

Measurement of the similarity of new case by the old case is done using the nearest neighbor for global similarity (global similarity). Meanwhile, to measure the level of local similarity between case bases and the case that will be predicted by each feature using euclidean distance for input features in the numerical form and hamming distance for non-numeric features that converted to binary. This calculating method of similarity calculates the closest distance between facts in a case that will be predicted with an existing case. This distance measurement focuses on values and measurements with greater values that show less similarity. After each feature is calculated the level of local similarity, later on its calculate the level of overall similarity (global similarity). Each feature has its own concern. The results of this global similarity determine how close the case base is to the case that will be predicted.

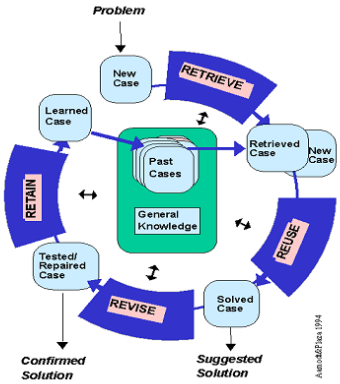


Figure 1 Case Based Reasoning Cycle

2. METHODS

2.1 System Description

This system was built to predict the time of graduation of students candidate of Computer Science at Gadjah Mada University who implemented the concept of case base reasoning (CBR). Basically CBR is one of the method that uses old experience solution to solve new problems. In general, the description of the system is shown in Figure 2. Input the system in the data form of master students candidate of the Computer Science (CS). The case prediction process is carried out by entering the data of the master students candidate of the Computer Science for new case.New cases that have been entered later processed to find similarities with the cases that stored in the case base. The process from the system will produce output in the form of a predicted student graduation time. [1]

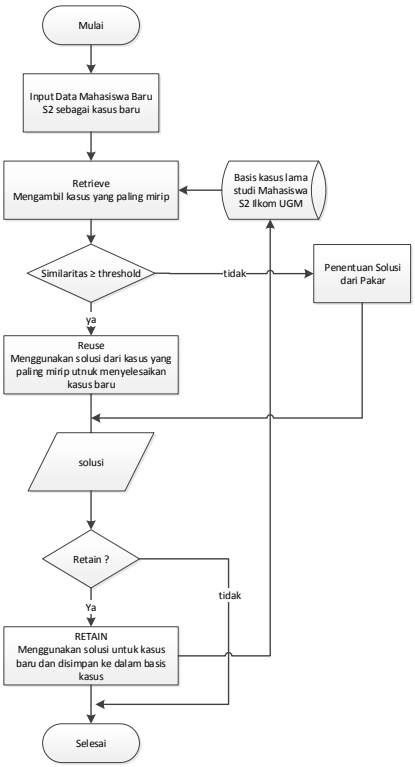


Figure 2 System Architecture

2.2 Euclidean Distance

Euclidean distance can be used to determine the distance between training data and testing data [7]. The euclidean distance formula [8] is shown in equation (1).

$$d_e(T,S^k) = \sqrt{\sum_{i=1}^{m} (T_i - S_i^k)^2} \quad (1)$$

Which is,

d_e = function distance between new problems and old cases
T = new problem
S^k = case base that exist in the storage
k = case base index, k = 1, 2, 3, ..., n
m = number of features in each case
i = feature index, i = 1, 2, 3, ..., m

2.3 Hamming Distance

Hamming distance can be defined as the number of bits that are different from the binary vectors that are compared. Hamming distance from two strings is the number of symbols from the two different strings. For example hamming distance between string "toned" and "roses" is 3. Hamming distance is also used to measure the distance between two binary strings, for example the distance between 10011101 and 10001001 is 2. The same things can be done to measure the proximity of binary numbers [9].

As for the formula to calculating hamming distance [8] is shown in equation (2).

$$d_h(T,S^k) = \sum_{i=1}^{m} |T_i - S_i^k| \quad (2)$$

Which is,

d_h = function of distance between new problems and old cases
T = new problem
S^k = case base that exist in the storage
k = case base index, k = 1, 2, 3, ..., n
m = number of features in each case
i = feature index, i = 1, 2, 3, ..., m

2.4 Nearest Neighbor

Nearest Neighbor is a method that uses the cumulative number of the feature weights which is suitable to the old case for the case that will be retrive.The nearest neighbor algorithm works by using a similarity pattern,so that the nearest neighbor uses the similarity calculation formula.

Similarity calculation aims to choose the most relevant or suitable case. The basic assumptions used are similar problems that will have similar solutions. As for the formula to calculating the proximity between two cases [11] is shown in equation (3).

$$Similarity\,(T,S^k) = \sum_{i=1}^{n} \frac{f(T_i,S_i^k) \times W_i}{W_i} \quad (3)$$

Which is,

T = new problem
S^k = case base that exist in the storage
k = case base index, k = 1, 2, 3, ..., n
m = number of features in each case
i = feature index, i = 1, 2, 3, ..., m
f = similarity function, problem T and case S^k
w_i = the weight of each feature

2.5 Standard Deviation

In research [10] which predicts emergency resource demand using CBR, weighting uses standard deviations to determine the weight of each feature. As for the formula to calculating the mean value of a feature is shown in equation (4).

$$\bar{m}(S^k) = \frac{1}{m}\sum_{i=1}^{m} S_i^k \quad (4)$$

Which is,

$\bar{m}(S^k)$ = average of case stored in the case base
m = the number of the features that stored in the case base
S_i^k = feature i in each case k
k = base case index, k = 1, 2, 3, ..., n
i = is a feature index, i = 1, 2, 3, ..., m

The formula for calculating the standard deviation is shown in equation (5).

$$\delta_i = \left[\frac{\sum_{i=1}^{m}\left(S_i^k - \bar{m}(S^k)\right)^2}{n}\right]^{\frac{1}{2}} \quad (5)$$

Which is,

δ_i = standard deviation of each feature
$\bar{m}(S^k)$ = average of the cases that stored in the case base
n = number of cases that stored in the case base
S_i^k = feature i in each case k
k = base case index, k = 1, 2, 3, ..., n

Based on the results of the calculation of the standard deviation, the weight of each feature can be obtained by the formula in equation (6).

$$w_i = \frac{\delta_i}{\sum_{i=1}^{m}\delta_i} \quad (6)$$

Which is,

w_i = the weight of each feature
δ_i = standard deviation of each feature
i = is a feature index, i = 1, 2, 3, ..., m
m = are many features in each case

2.6 Case Representation

CBR requires a knowledge base to obtain solutions, so its needed a representation of knowledge is needed as a knowledge base on the system. Knowledge on this system is represented by using a flat form. The knowledge that is meant is master students candidate data. Cases are represented in the form of a collection of the features that characterize the case and the solutions to handled. The representation includes the data of the Master's degree Computer Science students as a problem space and the graduating time for the Master's degree Computer Science students. The features used to make predictions are age, gender, occupation, distance traveled, scholarship, origin of the Bachelor degree universities, Bachelor degree college status, Bachelor degree study program, Bachelor degree study program accreditation, toefl score, Bachelor degree Grade Point Average (GPA) and the GPA's Master degree entry obtained by students candidate when following the new student admission test.

The type of work for students candidate is divided into several parts with different scores for example as shown in Table 1.

Table 1 Type of Work

Typo of Work	score
Government Employee Non-Lecturer	4
State Lecturer	3
Non-State Lecturer	2
Employees	1
Jobless	0

The origin of the Bachelor degree University is one of the features which states that students from Gadjah Mada University (UGM), Universitas Indonesia (UI), Institut Teknologi Bandung (ITB), and the Institut Teknologi 10 November (ITS) will receive direct study programs without tests. Based on the terms of the admission to the new student, the origin of the Bachelor Degree College is grouped into 6 groups, as shown in Table 2.

Table 2 Grouping Features From Bachelor Degree University

Group Name Bachelor Degree University	score
4 State University in same major	4
4 State University in different major	3
State University in same major	3
State University in different major	2
Non-State University in same major	2
Non-State University in different major	1

The Bachelor Degree Study Program states that the Bachelor Degree Study Program is grouped into 5 groups as shown in Table 3.

Table 3 Grouping of the features from the Bachelor Degree Study Program

Nama Kelompok Prodi S1	Score
Computer	5
Science	4
Technology	3
Other Science	2
Social Group	1

These features are then represented in the flats form and stored as the basis for the CBR system case as shown in Table 4.

Table 4 Base Case Representation

Domain: graduation time	
Case Number : 1	
Problem	
Case_Code	**0001**
Age	25
Gender	Female
Occupation	Non-State Lecturer
Mileage	15
Scholarship	Yes
Origin Bachelor Degree University	UGM
Bachelor Degree University Status	State
Pause Pass	2
Bachelor Degree Study Program	Computer Science
Bachelor Degree Study Program Acreditation	A
TOEFL Score	470
Bachelor Degree GPA	3.41
Master Degree Entry GPA	3.41
Length of Study	
Pass	22

2.7 Weighting

In this research will be calculated based on base case data using standard deviation methods which is suitable with research conducted by [10]. As for calculating the mean value of case data using equation (4), then calculating the standard deviation by using equation (5), and finally calculating the weight value using equation (5). Standard deviation data and weights for each feature are shown in Table 5.

Table 5 Standard deviation values and feature

Fitur	Standar Deviasi	Bobot Fitur
Usia	0,060637	0,035379
Jenis Kelamin	0,177065	0,103309
Pekerjaan	0,118523	0,069153
Jarak Tempuh	0,288358	0,168244
Beasiswa	0,179535	0,10475
PT S1	0,051761	0,0302
Status PT	0,155783	0,090892
Jeda lulus	0,088177	0,051447
Prodi S1	0,219651	0,128157
Akreditasi	0,178146	0,10394
Score Toefl	0,125658	0,073316
Ipk S1	0,03414	0,019919
Ipk Entry S2	0,036493	0,021292
Jumlahan	**1,713927**	**1**

2.8 Retrieve and Reuse

Retrieve used in this research is comparing and matching each new problem with the existing cases by calculating similarity. Similarity used to match feature values in a case is called local similarity. Whereas global similarity is used to find similarities between new cases that are targeted (T) and the old cases that become the source case (S).

Local similarity calculations are calculated by calculating the distance between new problems and cases in the case base. The smaller the distance between cases, the greater the level of similarity. To get the distance used euclidean formula distance and hamming distance. Euclidean distance can be calculated using equation (1) and hamming distance can be calculated using equation (2).

The features will be grouped into 2, that is to be calculated using euclidean distance and hamming distance. The feature that is calculated using euclidean distance is a feature whose input is in the form of numbers. Whereas for features that are calculated using hamming distance is a feature whose input is converted to binary that is 0 and 1 because there are only 2 choices.

Global similarity calculation in this research uses the nearest neighbor formula using equation (1). Target similarity measurements are made for all the source cases in the case base. The retrive process is shown by flowchat in Figure 3.

After all cases are matched, the next process find the highest global similarity value. The case with the highest similarity will be the solution for the process of adaptation to new prediction cases. This process is called the reuse process.

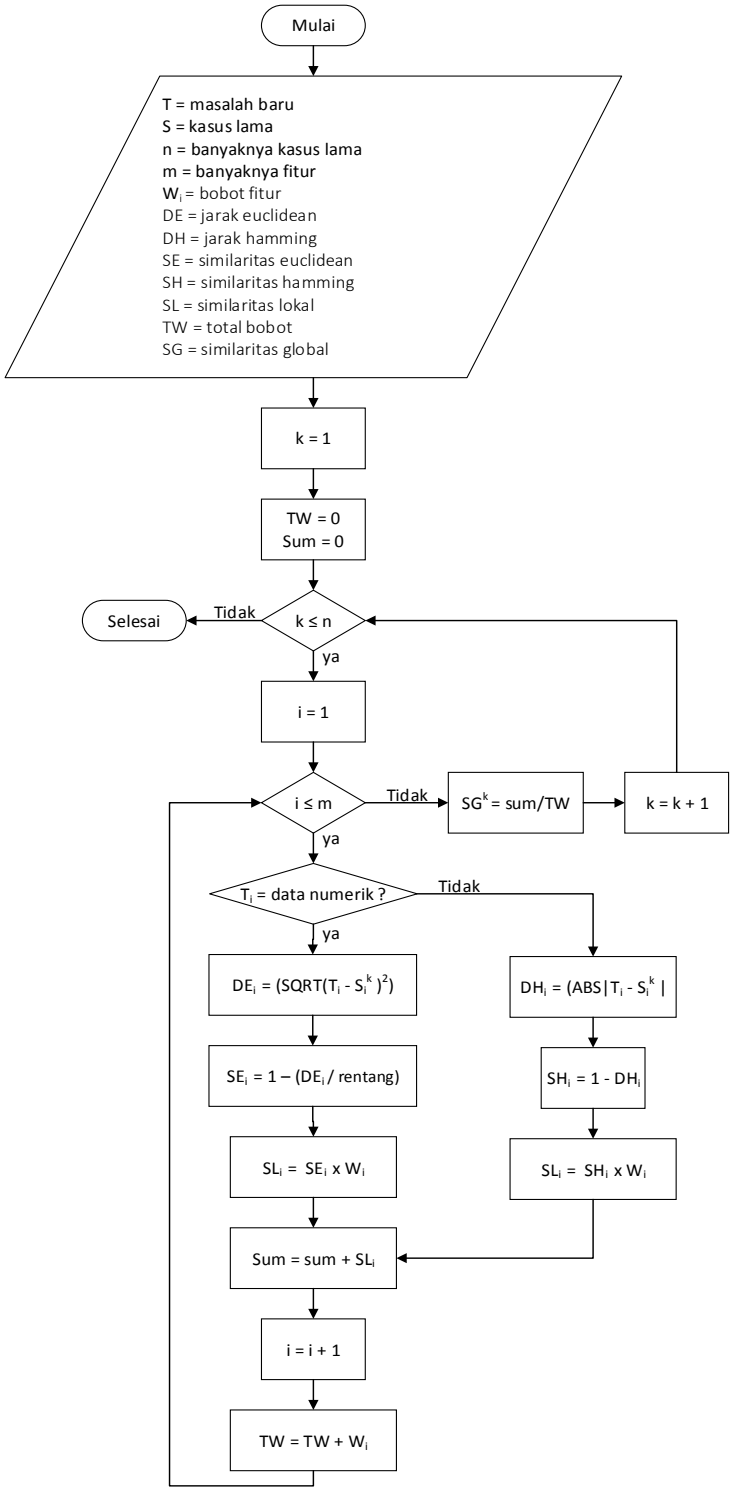


Figure 3 Retrieve Process Flowchart

2.8 Revise and Retain

The revise process is process of the case adaptation. In this research, this process will adapt if global similarity is less than the specified limit. The case that be predicted, will be stored in the prediction table to wait for the Head of Study Program revise the predict results. The Head of Study Program will revise the predictions results of the student study time in a

matter of months. After the case is revised, the case will be saved again in the prediction table.

The next process is the retain process. Where the storage process stores cases that are predicted to be the basic part of the case. This process will be done when the candidate student has passed the thesis examination.

2.9 Threshold

Threshold in this research is used to determine the threshold of global similarity values that are used to find the predicted solution. New cases that are predicted will be matched with the old cases thet stored in the case base. Case with global similarity values more than the threshold are candidates for the solutions. If there is no case of global similarity that is more than the threshold, then the case predicted will be revised by the Head of Study Program.

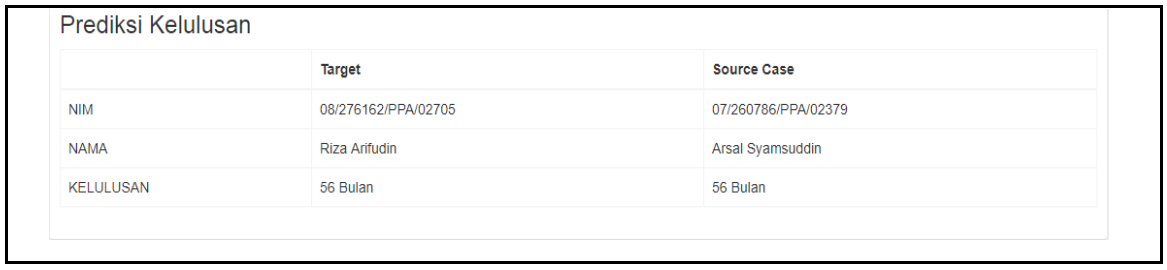
3. RESULTS AND DISCUSSION

3.1 Prediction Process

There are 4 main processes in Case Based Reasoning, there are retrieving, reuse, revise, and retain. The retrieval process is done when the Chief of Study Program do the prediction process in the system by input the Student ID or name of the student that will be predicted.

The retrieval process is done when the Chief of Study Program do the prediction process in the system by input the Student ID or name of the student that will be predicted. After the student that will be predicted has appeared, then the next step is the retrieving and reusing process.

The retrieval process is done out after the student that will be predicted appears, then the value features of the student will be displayed. At the retrieval stage there are 2 processes, there are the local similarity calculation process and the global similarity calculation process. After the system calculates the local similarity of the target with all sources, to calculate the global similarity, weights are needed. This reuse process is the process of taking the solution candidate from the highest global similarity case or the most similar case to a new case. The output of the prediction process is shown in Figure 4.


Prediksi Kelulusan

	Target	Source Case
NIM	08/276162/PPA/02705	07/260786/PPA/02379
NAMA	Riza Arifudin	Arsal Syamsuddin
KELULUSAN	56 Bulan	56 Bulan

Figure 4 Prediction Result

3.2 Test Result

The process of analyzing the prediction system ability for master students in Computer Science aims to determine the ability of the system in predicting new cases.Testing on the system is done by using accuracy testing to calculate the proximity or validity results identification of the actual data system. Tests were done using 50 test data and with 175 cases of data. The data that be tested will be predicted and then matched with the origin results.The prediction process is done as much as the test data with 38 results predicted to be correct. From these predictions results, then the accuracy is calculated using equation (7).

$$akurasi = \frac{38}{50} \times 100\% = 76\%$$

The calculation results above show the percentage of the system capability is 76%.

4. CONCLUSIONS

Based on the implementation, the Cased Based Reasoning system that was built using Euclidean Distance, Hamming Distance, and Nearest Neighbor can be applied to predict the graduation of Masters student candidate in Computer Science. The test results by calculating accuracy show that the system is able to predict the graduation of Master student of Computer Science by 76%.

The research that done is still limited to the data given by the Computer Science Study Program at Gadjah Mada University. The next research can be done by adding more specific features that can interfere the process of completing the master's degree and weighting using standard deviation methods so that further the research is recommended to use other weight calculation methods to get better weight results.

REFERENCES

[1] S. Salmu and A. Solichin, "Prediction of Timeliness Graduation of Students Using Naive Bayes: A Case Study at Islamic State University Syarif Hidayatullah Jakarta," *Prosiding Seminar Nasional Multidisiplin Ilmu,* pp. 701-709, 2017.

[2] X. Hu, B. Xia, M. Skitmore and Q. Chen, "The Application of Case-Based Reasoning in Construction Management Research: An Overview," *Automation in Construction,* no. 72, pp. 65-74, 2016.

[3] E. Wahyudi and S. Hartati, "Case-Based Reasoning untuk Diagnosis Penyakit Jantung," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems),* vol. 1, no. 11, pp. 1-10, 2017.

[4] T. Rismawan and S. Hartati, "Case-Based Reasoning untuk Diagnosa Penyakit THT (Telinga Hidung dan Tenggorokan)," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems),* vol. 6, no. 2, pp. 67-68, 2013.

[5] S. Kosasi, "Penerapan Metodologi Penalaran Berbasis Kasus dalam Mendiagnosa Kerusakan Komputer," *TECHSI,* vol. 7, no. 2, pp. 187-202, 2015.

[6] M. T. Rezvan, A. Z. Hamadani and A. Shalbafzadeh, "Case-Based Reasoning for Classification in the Mixed Data Sets Employing the Compound Distance Methods," *Engineering Applications of Artificial Intelligence,* vol. 26, no. 9, pp. 2001-2009, 2013.

[7] N. Krisandi, Helmi and B. Prihandono, "Algoritma k-Nearest Neighbor dalam Klasifikasi Data Hasil Produksi Kelapa Sawit pada PT. Minamas Kecamatan Parindu," *Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster),* vol. 2, no. 1, pp. 33-38, 2013.

[8] K. J. Cios, P. Witold, R. W. Swiniarski and L. A. Kurgan, Data Mining: A Knowledge Discovery Approach, Springer, 2007.

[9] L. S. Putro, R. Saptono and R. Anggrainingsih, "Penerapan Kombinasi Algoritma Minhash dan BInary Hamming Distance pada Hybrid Perekomendasi LAgu," *Jurnal ITSMART,* vol. 2, no. 1, pp. 36-43, 2013.

[10] W. Liu, G. Hu and J. Li, "Emergency Resources Demand Prediction using Case-Based Reasoning," *Safety Science,* vol. 50, no. 3, pp. 530-534, 2017.

[11] E. Wahyudi, Case-Based Reasoning untuk Diagnosis Penyakit Jantung, Yogyakarta: Gadjah Mada University, 2015.